

Spatio-Temporal Image-Based Encoded Atlases for EEG Emotion Recognition

Danilo Avola ^{*}, Luigi Cinque [†], Angelo Di Mambro [‡], Alessio Fagioli [§],
Marco Raoul Marini  and Daniele Pannone [¶]

*Department of Computer Science, Sapienza University of Rome
Via Salaria 113, Rome 00198, Italy*


^{*}*avola@di.uniroma1.it*

[†]*cinque@di.uniroma1.it*


[‡]*dimambro@di.uniroma1.it*

[§]*fagioli@di.uniroma1.it*

[¶]*pannone@di.uniroma1.it*

Bruno Fanini 

*Institute of Heritage Science, National Research Council
Area della Ricerca Roma 1, SP35d, 9, Montelibretti 00010, Italy
bruno.fanini@cnr.it*

Gian Luca Foresti 

*Department of Computer Science, Mathematics and Physics
University of Udine, Via delle Scienze 206, Udine 33100, Italy
gianluca.foresti@uniud.it*

Received 30 September 2023

Accepted 9 February 2024

Published Online 27 March 2024

Emotion recognition plays an essential role in human–human interaction since it is a key to understanding the emotional states and reactions of human beings when they are subject to events and engagements in everyday life. Moving towards human–computer interaction, the study of emotions becomes fundamental because it is at the basis of the design of advanced systems to support a broad spectrum of application areas, including forensic, rehabilitative, educational, and many others. An effective method for discriminating emotions is based on ElectroEncephaloGraphy (EEG) data analysis, which is used as input for classification systems. Collecting brain signals on several channels and for a wide range of emotions produces cumbersome datasets that are hard to manage, transmit, and use in varied applications. In this context, the paper introduces the Empátheia system, which explores a different EEG representation by encoding EEG signals into images prior to their classification. In particular, the proposed system extracts spatio-temporal image encodings, or atlases, from EEG data through the Processing and transfer of Interaction States and Mappings through Image-based eNcoding (PRISMIN) framework, thus obtaining a compact representation of the input signals. The atlases are then classified through the Empátheia architecture, which comprises branches based on convolutional, recurrent, and transformer models designed and tuned to capture the spatial and temporal aspects

*Corresponding author.

This is an Open Access article published by World Scientific Publishing Company. It is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND) License which permits use, distribution and reproduction, provided that the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

of emotions. Extensive experiments were conducted on the Shanghai Jiao Tong University (SJTU) Emotion EEG Dataset (SEED) public dataset, where the proposed system significantly reduced its size while retaining high performance. The results obtained highlight the effectiveness of the proposed approach and suggest new avenues for data representation in emotion recognition from EEG signals.

Keywords: Emotion recognition; image encoding; spatio-temporal atlases; multi-branch architecture; EEG; PRISMIN framework; CNN; LSTM; GRU; ViT.

1. Introduction

Emotions are a key human trait that can be defined as a biological state associated with neurophysiological changes that are linked with thoughts, feelings, behavioral responses, pleasure, or displeasure sensations¹ and can affect almost every aspect of our existence, such as, among others, social interactions, relational life, work productivity, and even human-computer interactions. Although different communication channels can be used to express emotions, e.g. facial expressions, voice pitch, and posture, it can be difficult to understand these cues that convey additional information about a person.² As a consequence, emotion recognition through automatic approaches can be extremely useful in diverse relevant scenarios. A first example regards the diagnosis of depressive states or post-traumatic stress disorders (PTSD)³ to identify if a patient is experiencing pain during a treatment,⁴ another interesting case study concerns the diagnosis of Parkinson's Disease (PD)⁵ to understand if a patient exhibits emotional impairments when emotionally elicited, a last example is the detection of fake emotions during an interrogation in court.⁶ Actually, nowadays, this type of technology finds interest in an increasingly wide range of application fields. As a matter of fact, many works in the current literature try to address the emotion recognition task by exploiting facial traits,^{7,8} body movements,^{9,10} speech,¹¹ multimodal approaches,^{12–14} or even more complex data such as brain electrical activity.^{15–17} Concerning the latter, it can be measured through ElectroEncephaloGraphy (EEG), which extracts brainwaves through the use of surface electrodes. Five different waves are retrieved with the EEG, i.e. delta, $\delta \in [1.5–4 \text{ Hz}]$,¹⁸ theta, $\theta \in [5–8 \text{ Hz}]$,¹⁹ alpha, $\alpha \in [9–14 \text{ Hz}]$,²⁰ beta, $\beta \in [15–40 \text{ Hz}]$,²¹ and gamma, $\gamma \in [25–140 \text{ Hz}]$.²² These waves respond to specific activities, including daydreaming or active thought, via working memory

and attention. What is more, brainwaves are also related to emotion processing, where specific patterns in high-frequency bands are associated with positive, neutral, and negative feelings through time-frequency analysis.²³ As can be observed in Fig. 1, positive emotions show an increment in energy for beta and gamma frequency bands, whereas neutral and negative emotions present decreased beta and gamma energy. While the neural patterns of negative and neutral emotions have similar patterns in beta and gamma bands, the latter have higher energy of alpha oscillations. Although in this work, we utilize a collection of EEG signals previously acquired by another research group, it remains important to explain how these databases are typically created. This is also to introduce the motivations behind the approach proposed in this paper.

Brainwaves are commonly acquired through Brain-Computer Interfaces (BCIs), devices that

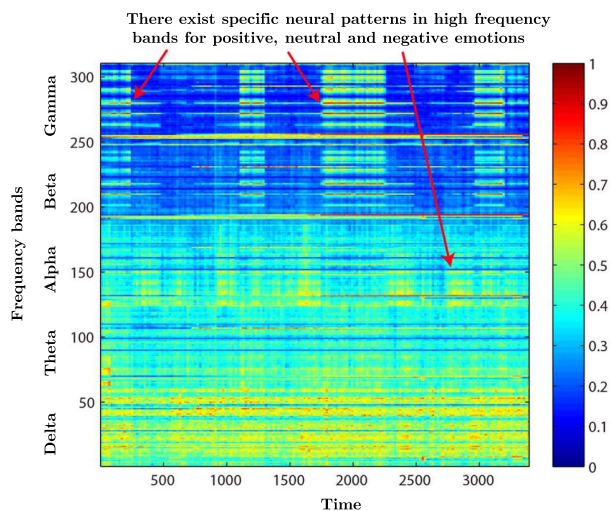


Fig. 1. Emotions captured via EEG signals. The differential entropy (DE) energy feature (scaled in $[0,1]$) is represented for each frequency band (in Hz) over time (in ms), highlighting specific patterns in higher frequency ranges corresponding to specific emotions.

enable the capture of brain signals through a set of electrodes. Depending on their proximity to the brain tissue, these electrodes can be invasive or noninvasive. Specifically, invasive electrodes, which are more accurate, require neurosurgery to implant them directly into the brain tissue. On the other hand, noninvasive electrodes are less accurate, as they are based on an external EEG helmet positioned on the subject's scalp, but users more commonly accept them. In addition, recent advances in the accuracy of noninvasive electrodes make these interfaces comparable to invasive ones in several practical contexts.^{24–26} Anyway, regardless of the electrode typology, BCIs are being used by a wide audience in many applications, including emotion recognition.²⁷ More specifically, these devices are used to acquire extensive data for training classifiers. Unlike other types of digital information, collecting EEG data can quickly become complex and cumbersome, depending on the number of channels and the range of emotions that need to be monitored. As for other digital resources, compression techniques can support several critical aspects, including the reduction of storage space, which assumes particular importance in contexts where storage is expensive or limited, such as in mobile devices, drones, robots, and in general, embedded systems; transmission efficiency, which becomes crucial in environments where bandwidth is bounded or communication channels are constrained, such as in long-distance monitoring; compatibility and scalability, where data compression allows us for easy scaling or conversion of data to different formats that can be compatible with various devices, applications, or protocols; among other advantages.

To address the issues reported above, this paper introduces the novel Empátheia^a system, which encodes EEG data into images, referred to as atlases in this work, before classifying the underlying emotion. In detail, the proposed approach pre-processes multichannel EEG signals and generates spatio-temporal atlases using an encoder based on Processing and transfer of Interaction States and Mappings through Image-based eNcoding (PRISMIN) framework.²⁸ This framework compresses brain signals into a coarse visual representation, i.e. an image. Then, the system uses a deep learning-based pipeline

as a classifier that recognizes the emotion. Specifically, the architecture is composed of branches based on convolutional, recurrent, and transformer models designed and tuned to capture the spatial and temporal aspects of an emotion represented by the atlas. Extensive experiments were conducted for both the encoder and classifier. In particular, two encoding types, i.e. short-rainbow (SR) and grayscale (GS), and four different models, i.e. one-based exclusively on a convolution neural network (CNN), two based on mixtures of CNN and RNN, and a last based on a transformer, were tested to find an effective emotion recognition method. The Empátheia system was evaluated on the Shanghai Jiao Tong University (SJTU) Emotion EEG Dataset (SEED).^{23,29} The proposed approach significantly reduced the dataset size, thus enabling the implementation of less computationally intensive models. In addition, it allows us a faster training while retaining high performance on the emotion recognition task. What is reported highlights the effectiveness of the proposed method and, at the same time, suggests new pathways for EEG signal management and processing. The main contributions of this work can be summarized as follows:

- A novel spatio-temporal atlas representation for EEG data thanks to a custom encoder based on the PRISMIN framework.
- The development of the Empátheia system, a multi-branch classifier made up by four independent classifiers, each one designed to capture spatial and temporal features of an image-encoded emotion.
- An alternative way to treat and manage large collections of EEG signals to support several critical issues, including storage space, embedded systems, transmission efficiency, and many others.
- Exploring different encoding strategies and classifiers to retain high performance on the SEED benchmark dataset despite the significant data quantization produced by the PRISMIN encoder.

The remainder of this paper is structured as follows. Section 2 presents an overview of related work addressing the emotion recognition task. Section 3 describes the Empátheia system, providing details on

^aFrom ancient Greek *en-*, “inside”, and *pathos*, “sentiment, feeling”. Today is translated as “Empathy”.

the PRISMIN encoder and Empátheia classifier. Section 4 introduces the SEED public dataset and discusses the results obtained via different encoding strategies and classifiers, as well as a comparison with the current state-of-the-art. Finally, Sec. 5 draws a conclusion to the presented work.

2. Related Work

An important aspect of EEG emotion recognition is the feature extraction from brain signals, which can affect classification accuracy. After signal pre-processing, for instance, downsampling³⁰ and band-pass frequency filtering,³¹ EEG features can be divided into single-channel and multichannel features. The former class was generally the most common choice in earlier works due to its proven effectiveness,³² and includes, among others, Power Spectral Density (PSD),³³ DE,²⁹ and Wavelet Features.³⁴ The second class has instead become the preferred option in recent years, especially with the evolution of deep learning approaches for EEG emotion recognition, resulting in various solutions available in the literature exploiting CNN, RNN, or graph-based architectures.³⁵ Although single-channel and multichannel features can be employed with deep learning approaches, the feature extraction in this work differs from these methods. Specifically, a multichannel atlas, i.e. an image associated with an emotion, is used to extract features automatically using a deep learning model.

Regarding CNN-based approaches, they tend to focus on spatial information derived from multiple EEG channels. Liu *et al.*,³⁶ for instance, apply a Butterworth band-pass filter to the EEG channels and reorganize the data to be used in a custom deep neural network (DNN) composed of a CNN, a sparse autoencoder (SAE), and a DNN, which are trained separately to enhance convergence speed. Instead, Liu *et al.*³⁷ utilize an attention mechanism to integrate spatial information into input signals and employ both a pre-trained convolutional capsule network for feature extraction and a secondary double-layer capsule network. A different example is discussed in Li *et al.*,³⁸ which focuses on the frontal lobe using Papez circuit theory. The authors utilize a frontal lobe double dueling DQN (FLD3QN) procedure based on reinforcement learning with EEG

channels and a bifrontal lobe residual CNN (BiFRCNN) for emotion recognition. Miao *et al.*³⁹ present a 3D deep residual learning framework for analyzing EEG signals across multiple frequency bands. They use group sparse regression for optimal frequency band selection and a 3D deep residual network for feature classification. Finally, Hu *et al.*⁴⁰ introduce a scaling layer in their convolutional network to extract spectrogram-like features from EEG signals. This layer uses varied convolutional kernels to identify patterns across different scales, eliminating the need for other feature extraction methods like DE.

Concerning RNN-based methods, they tend to focus on spatial and temporal information, which can be naturally captured by recurrent architectures. For example, Li *et al.*⁴¹ investigate the emotional expression differences between the brain's left and right hemispheres, using four RNNs across two brain regions to analyze spatial relationships. They design a subnetwork to integrate these hemispheres, thus enhancing emotion recognition feature extraction. Similarly, Zhang *et al.*⁴² utilize a multidirectional RNN to analyze long-range contextual cues in EEG data for capturing spatial variations in human emotions. They use projection matrices on spatial and temporal states to identify emotion-rich regions. The study described by Guo *et al.*⁴³ leverages the domain adaptation (DA) concept and aims to reduce the inter-session variability of EEG signals by designing a spatio-temporal feature extractor. The extracted features are then aligned to classify emotions. Yang *et al.*⁴⁴ use a combination of LSTM and CNN networks to analyze spatio-temporal features in raw EEG signals. The LSTM captures contextual data, while the CNN identifies inter-channel correlations via a 2D signal representation. Finally, Du *et al.*⁴⁵ develop an attention-based LSTM with domain discriminator (ATDD-LSTM) for spatial feature extraction across EEG channels. This approach focuses on nonlinear relations among electrodes to optimize EEG channel selection and minimize feature discrepancies across different subjects and sessions.

In relation to graph-based approaches, they exploit the natural configuration of BCI devices to model graph-like architectures. For instance, Liu *et al.*⁴⁶ introduce a global-to-local feature aggregation network (GLFANet) that uses topological graphs to analyze spatial and frequency domain

features in EEG channels. The network employs both global (graph convolutional blocks) and local (convolutional blocks) learners to extract EEG signal features. The work described by Zhong *et al.*⁴⁷ focuses on left and right hemisphere coupling in emotion recognition using a regularized graph neural network (RGNN). They analyze local and global EEG channel relations, finding pre-frontal, parietal, and occipital regions notably informative. Song *et al.*⁴⁸ use a dynamic graph convolution neural network (DGCNN) to model multichannel EEG features, learning an adjacency matrix during training to represent relationships among EEG channels for feature discrimination. Differently, Zhou *et al.*⁴⁹ introduce a progressive graph convolution network (PGCN) for identifying coarse and fine-grained emotional features. Their dual-graph module encapsulates dynamic functional connections and static spatial brain region data. Finally, Yin *et al.*⁵⁰ present a system combining graph convolutional neural network (GCNN) and LSTM network to analyze EEG signals. The GCNNs generate domain features from DE processed signal segments, and the LSTMs then extract temporal features and classify emotions by channel relationship.

3. Proposed Method

The Empátheia system addresses the EEG emotion recognition task using a reduced amount of data. It can be divided into two modules: PRISMIN atlas encoder and Empátheia classifier. The first module, discussed in Sec. 3.1, compresses EEG signals into atlases. These atlases serve as coarse visual representations of the EEG signals. The second module, presented in Sec. 3.2, performs the classification of the generated atlases, effectively achieving EEG emotion recognition. The architecture of the Empátheia system is reported in Fig. 2.

3.1. PRISMIN encoder

The first step, typically adopted by many works in the literature, involves direct EEG data pre-processing or feature extraction to remove signal artifacts, e.g. ocular artifacts, or extract specific features from the original data, e.g. extracting specific frequency bands. In contrast, the Empátheia system takes a different approach, conducting a pre-processing

phase using the PRISMIN framework to transform raw EEG signals into coarse spatio-temporal image atlases that describe emotions. In particular, the open-source PRISMIN framework²⁸ encodes data, such as attributes and user states, into compact and lightweight 2D images, making them easy to manipulate and transfer. The framework offers methods to encode session data into image atlases, including runtime accessories that can be employed in interactive sessions to capture and encode specific attributes. Regarding signal compression, as detailed in Fanini and Cinque,²⁸ the temporal layout combined with lossless image formats in scenarios where smooth variations of neighboring pixels are present⁵¹ produce optimal compression ratios. This is also performed by maintaining computationally simple and fast encoding/decoding routines that may operate on both CPUs or GPUs.

In the presented work, the main focus is to design an encoder based on PRISMIN that can transform large EEG datasets into compact, time-driven image atlases.⁵² To achieve this, PRISMIN allows to define the prism class, i.e. a custom data encoder. Specifically, a given prism \mathcal{P} provides:

- A *refract* method to define how incoming data are encoded as well as the layout adopted in the final image atlas. In this work, we use spatio-temporal atlases, although different methods can be used to implement the refract method.
- A *bake* method to write the actual atlas on disk, using specific image format and bit-depth. This can be implemented as on-demand routine, for instance, to control write accesses on server-side storages within service-based deployments of PRISMIN framework.

To adapt \mathcal{P} to handle EEG signals, they must be described so that the *refract* method can accept them. Intuitively, EEG records are represented by streams of voltage values captured by the C channels of a BCI device over a given time period T for a specific EEG session. Formally, a session S containing the brain signals can be defined via the following matrix:

$$S(c, t) = v, \quad (1)$$

where $v \in \mathbb{R}$ corresponds to the voltage value of a given channel $c \in C$ at a given time instant $t \in T$. With this formalization, the session matrix S can be

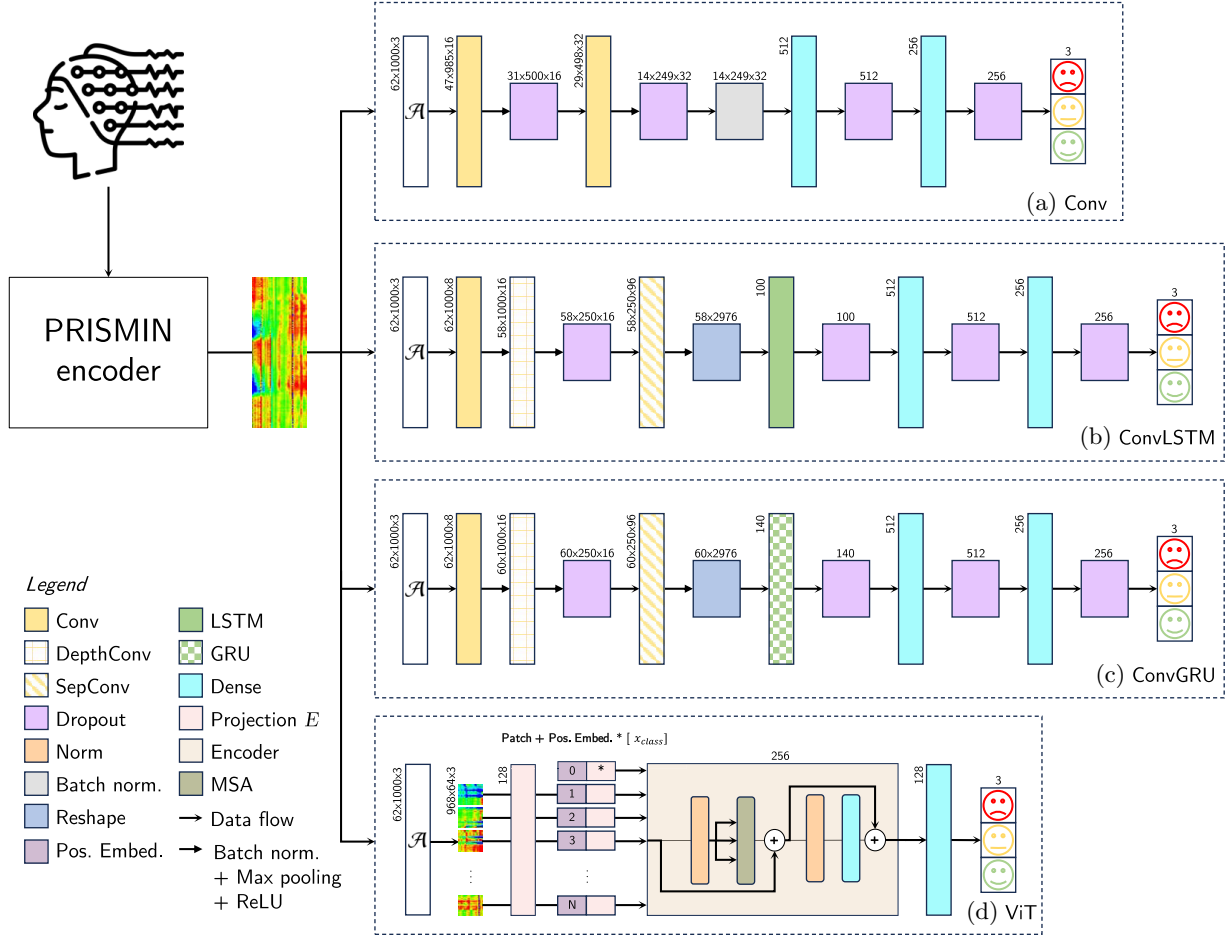


Fig. 2. Architecture overview of the Empátheia system. The PRISM encoder transforms EEG signals associated with emotions into 2D atlases, which are then utilized by the Empátheia classifiers for emotion recognition.

used to define a time-driven layout for the generated atlases, where rows and columns of the encoded image represent, respectively, the channels and time frames of the EEG signals.

A quantization error is indeed introduced by *refraction* of voltages (values v) on both SR and GS. Quantization error in PRISM depends on this specific scenario on: (1) voltage ranges, (2) color space adopted, and (3) image bit-depth. An in-depth analysis is described in Fanini and Cinque.^{28,52} Specifically, given Δ_v as voltage range and bit-depth b to encode incoming values, the maximum quantization errors for GS (ϵ_{GS}) and SR (ϵ_{SR}) are given by

$$\epsilon_{GS} = \pm \frac{\Delta_v}{2^{b+1}}; \quad \epsilon_{SR} = \pm \frac{\Delta_v}{2^{b+3}}. \quad (2)$$

After defining the *refract* and *bake* methods, a quantized voltage session prism \mathcal{P}_q can be initialized

to encode EEG signals into atlases. By applying \mathcal{P}_q over the EEG dataset, the atlases $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n)$ can be computed, where n corresponds to the number of samples in the EEG collection. Examples of \mathcal{P}_q -generated atlases using two different color spaces are shown in Fig. 3.

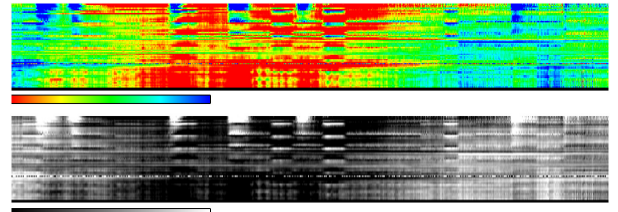


Fig. 3. (Color online) Examples of \mathcal{P}_q -generated atlases using (top) a SR color space and (bottom) GS. Each row represents a different EEG channel, pixel color is the encoded voltage over time (x -axis).

3.2. Empátheia classifier

To perform EEG emotion recognition on the encoded atlases \mathcal{A} , it is necessary to implement a classifier. Given the novelty of the approach and the 2D image representation of \mathcal{A} , a natural choice for classification would be the implementation of a CNN model. However, the atlases encode spatio-temporal characteristics of a given emotion, suggesting that other architectures, such as RNN-based or transformer-based approaches, could also be effective. In this context, several models that can serve as Empátheia classifiers are described in the following sections.

3.2.1. Conv classifier

The first classifier, depicted in Fig. 2(a), is a simple CNN architecture since the encoded EEG signals are effectively transformed into images. Specifically, it consists of two convolutional layers, each followed by batch normalization and max pooling operations, as well as the ReLU activation function, which is defined as follows:

$$f(x) = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

The peculiarity of these convolutions lies in their kernel size, which is large enough, e.g. 16×16 , to capture temporal aspects of the atlas. Moreover, these layers are responsible for extracting feature maps from the input atlases, which are then classified, after a flattening operation, using three dense layers. The latter comprises fully connected layers interleaved with dropout layers to enhance the abstraction capabilities of the model. Furthermore, the first two dense layers employ a ReLU activation function, while the last one uses a softmax function to compute the probability distribution over the available emotions via:

$$\sigma(x)_i = \frac{e^{x_i}}{\sum_{j=1}^K e^{x_j}}, \quad (4)$$

where x_i indicates the i th class score, K corresponds to the number of classes, i.e. the emotions, while x_j is used to normalize the obtained score over all the available classes. Finally, the model is trained using the categorical cross-entropy loss, which is defined as

follows:

$$\mathcal{L}_{\text{CE}} = - \sum_i^K y_i \log \hat{y}_i, \quad (5)$$

where y_i and \hat{y}_i correspond to the ground truth and predicted class probability, respectively, while K is the number of classes.

3.2.2. ConvLSTM classifier

The second classifier, shown in Fig. 2(b), takes inspiration from methods that combine CNN and RNN models to classify spatio-temporal characteristics of EEG signals.^{43,44} Differently from the first model, the ConvLSTM uses three convolutions, which include standard, depth, and separable convolutions. Additionally, to fully exploit the temporal aspect encoded in the atlases \mathcal{A} , the feature maps generated by these layers are analyzed through a bidirectional LSTM layer.⁵³ This LSTM implements forward and backward layers that inspect the provided input in both directions. Regardless of the data flow direction, an LSTM contains memory cells with input, forget, and output gates, as well as cell and hidden states. Formally, the LSTM at a given time step t and the previous hidden state h_{t-1} is defined as

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \quad (6)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \quad (7)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \quad (8)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t, \quad (9)$$

$$h_t = o_t \odot \tanh(c_t), \quad (10)$$

where σ is the sigmoid activation function; W_i , W_f , W_o , and b_i , b_f , b_o indicate the weight matrices and bias terms for the input, forget, and output gates, respectively; $[h_{t-1}, x_t]$ is the concatenation of the previous hidden state h_{t-1} and the input at time step t ; \odot denotes element-wise multiplication; c_t and \hat{c}_t correspond to the updated and candidate cell state at time step t , respectively; while \tanh is the hyperbolic tangent activation function. Finally, the model is trained using the same categorical cross-entropy loss described in Eq. (5).

3.2.3. ConvGRU classifier

The third classifier, represented in Fig. 2(c), follows the same structure as the ConvLSTM introduced in

Sec. 3.2.2 and is trained using the categorical cross-entropy loss presented in Eq. (5). However, instead of a bidirectional LSTM layer, ConvGRU exploits the gated recurrent unit (GRU), another type of recurrent neural network that can handle temporal data but has fewer parameters than the LSTM and can often perform similarly. In particular, the GRU has simpler memory cells with update and reset gates, as well as candidate and hidden states, that are formally defined as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z), \quad (11)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r), \quad (12)$$

$$\hat{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t]) + b_h, \quad (13)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t, \quad (14)$$

where σ is the sigmoid activation function; W_z , W_r , W_h , and b_z , b_r , b_h are the weight matrices and bias terms for the update gate, reset gate, and candidate cell state, respectively; $[h_{t-1}, x_t]$ is the concatenation of the previous hidden state h_{t-1} and the input at time step t ; \odot denotes element-wise multiplication; while \tanh represents the hyperbolic tangent activation function.

3.2.4. Vision transformer

The last classifier, shown in Fig. 2(d), is a fine-tuned version of the Vision Transformer (ViT).⁵⁴ The ViT model, inspired by the NLP transformer, splits the input images into patches to provide a sequence of linear embeddings given as input to a transformer, the same way as tokens in NLP applications. Starting from the encoded atlas \mathcal{A} as a 2D image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, it is handled into a sequence of flattened squared patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where H, W are the image height and width, respectively, C represents the number of channels, P is the dimension of each patch, and $N = \frac{HW}{P^2}$ is the total number of patches. To embed each patch into the model dimension D , a trainable linear projection E is applied, thus obtaining an embedding sequence \mathbf{z}_0 , which is defined as

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1 \mathbf{E}; \mathbf{x}_p^2 \mathbf{E}; \dots; \mathbf{x}_p^N \mathbf{E}] + \mathbf{E}_{\text{pos}}, \quad (15)$$

where $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$ and $E_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$. Furthermore, a learnable 1D position embedding is added to each patch embedding to retain positional information, then a learnable class embedding ($\mathbf{z}_0^0 = \mathbf{x}_{\text{class}}$) is

prepended to the patches sequence, representing the image label \mathbf{y} (Eq. (18)) at the output (\mathbf{z}_L^0) of the L -layers transformer encoder. The resulting sequence of embedding vectors is the input of the transformer. The transformer uses a constant latent vector of size D and is composed of alternating layers of Multi-headed Self-Attention (MSA)⁵⁵ and a two-layer perceptron, both preceded by a LayerNorm (LN) layer and followed by a residual connection. Finally, a classification head, implemented as a linear layer, is attached to \mathbf{z}_L^0 . Formally, the transformer blocks are defined as follows:

$$\mathbf{z}_t = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \quad (16)$$

$$\mathbf{z}'_{\ell} = \text{MLP}(\text{LN}(\mathbf{z}'_{\ell})) + \mathbf{z}_{\ell}, \quad (17)$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0), \quad (18)$$

where $\ell = 1 \dots L$. Finally, the model is trained using the same categorical cross-entropy loss described in Eq. (5).

4. Experimental Results

This section assesses the effectiveness of the proposed approach in EEG emotion recognition. In detail, Sec. 4.1 introduces the public dataset used to evaluate the Empátheia system. Section 4.2 reports the implementation details required to reproduce the experiments. Section 4.3 examines the performance of the PRISMIN atlas encoder. Finally, Sec. 4.4 evaluates the Empátheia classifier through ablation studies and a state-of-the-art comparison.

4.1. Dataset

The dataset used to train and test the Empátheia system is the SEED,²³ a public collection focusing on the EEG emotion recognition task. The dataset is composed of 15 subjects, 7 males, and 8 females, all right-handed students of Shanghai Jiao Tong University. A total of 6 clips have been shown to the participants of the experiment, and each clip is associated with negative, neutral, and positive emotions. Finally, each emotion has 5 corresponding emotional clips. Each trial comprehends a 5 s hint before each clip, a roughly four-minute-long clip, followed by 45 s for self-assessment, and is concluded with 15 s of rest. The dataset is provided already pre-processed by the authors. In detail, the original EEG

data were downsampled to a sampling rate of 200 Hz. Then, visual inspection of the data was performed on the new EEG signals, and recordings significantly affected by electromyographic (EMG) and electro-oculographic (EOG) interferences were manually excluded. EOG data, recorded during the experiments, were also utilized to identify blink artifacts within the recorded EEG data. To mitigate noise and remove artifacts, the EEG data have been processed using a bandpass filter with a range of 0.3–50 Hz. Subsequently, to the preprocessing process, EEG segments corresponding to the duration of each movie were extracted. Each channel of the EEG data was then divided into nonoverlapping epochs of equal length, i.e. 1 s. For a single experiment, approximately 3300 clean epochs were obtained.

The detailed acquisition protocol for a single trial is summarized in Fig. 4.

Regarding the acquisitions, EEG signals were collected using the 62-channel ESI NeuroScan System at a sampling rate of 1000 Hz, according to the international 10–20 system for 62 channels. Furthermore, the authors down-sampled the signals to 200 Hz and applied a 0–75 Hz bandpass frequency filter. The resulting EEG signal segments correspond to the duration of each clip; therefore, their length can slightly differ when considering distinct segments.

4.2. Implementation details

The Empátheia system was implemented using the PyTorch framework.^b All experiments were performed using 80/10/10% splits for training, validation, and test sets, respectively. All models were trained using the AdamW optimizer for 100 epochs using the same hyper-parameters, i.e. $1e-03$ learning rate and a batch size of 64. Standard classification

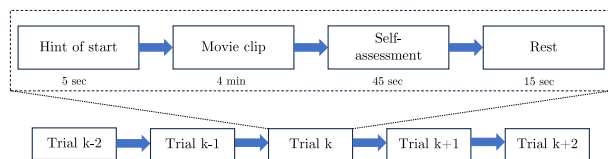


Fig. 4. Detailed protocol used during experiments on SEED dataset.

metrics, i.e. accuracy, precision, recall, and F1-score, were employed to assess the system.

The experiments were executed using an AMD EPYC 7301 16-Core Processor with 64 GB of RAM and an RTX QUADRO 6000 with 24 GB of RAM.

4.3. PRISMIN encoder evaluation

The first component to be evaluated is the PRISMIN encoder tasked with the atlas generation. In particular, since the primary focus of the encoder is to reduce the SEED dataset size, its assessment revolves around the compression rate (C-Rate) of the input dataset. To achieve this goal, this paper explores the effectiveness of two different encodings, i.e. linear GS and SR mappings, during the implementation of the *refract* and *bake* methods discussed in Sec. 3.1. The resulting compressed datasets are reported in Table 1. In detail, for each sample trial in the SEED collection, the devised PRISMIN encoder generates a PNG atlas containing either its GS or SR encoding. The resulting atlas represents the entire EEG trial and has a final shape of $62 \times 10,800$, corresponding to the number of channels and recording length of the captured EEG signals. Note that the width of the generated atlases depends on the corresponding session length. Therefore, to ensure that all image-encoded signals have the aforementioned shape, the atlases are cropped on the left and right sides by up to 10% of their width. These portions are generally less relevant for emotion recognition as they are associated with the start and end of an EEG acquisition. From this procedure, the PRISMIN encoder effectively constructs two new datasets, i.e. \mathcal{D}_{GS} and \mathcal{D}_{SR} , corresponding, respectively, to the GS and SR encodings, that both maintain the number of samples of the original SEED dataset, i.e. 675. However, due to the performed compression, the resulting collections manage

Table 1. Dataset reduction using PRISMIN encoding.

\mathcal{D}	Baseline	C-Rate	Augmented	C-Rate
\mathcal{D}_{GS}	998 MB	6.9×	744 MB	9.2×
\mathcal{D}_{SR}	670 MB	10.3×	907 MB	7.6×

^bSource code is available at: https://github.com/Prometheus-Laboratory/2024_prismin.

to significantly reduce the original dataset size by a factor of $10.3\times$ and $6.9\times$, attesting to a disk space of 670 MB and 998 MB for the GS and SR encoding, respectively.

The direct conversion of the SEED dataset using the described PRISMIN encoder effectively reduces the collection size. Despite that, the generated atlases depict a coarse representation of the original EEG signals and do not allow deep learning models to learn the emotion recognition task. What is more, even works analyzing EEG signals from the SEED dataset tend to suffer from this issue, which is generally addressed by applying a data augmentation strategy through slicing of the original recordings.^{36,56} Thus, following this rationale, starting from the beginning of the generated atlases, they are sequentially split into smaller ones by extracting sub-frames of shape 62×1000 , as depicted in Fig. 5, with a 200 pixels overlap on the x -axis among subsequent sub-frames. This approach, already used in different literature works,^{36,39,56} is adopted to partially preserve middle information among subsequent frames, limiting emotions cut-off. With this configuration, the resulting \mathcal{D}_{GS} and \mathcal{D}_{SR} datasets, used to evaluate the Empátheia classifiers, contain 9450 samples instead of 675 and enable the implemented models to perform emotion recognition, as reported in Sec. 4.4.

Even after the reported data augmentation strategy, \mathcal{D}_{GS} and \mathcal{D}_{SR} still considerably reduce the original SEED dataset size by factors of $9.2\times$ and $7.6\times$ for the GS and SR encodings, respectively. Interestingly, for the GS encoding, when considering smaller image portions, the PRISMIN encoder further compresses the atlases, resulting in an even smaller collection. This outcome is possibly due to the reduced amount of noise in the GS sub-frames and suggests that additional encoding strategies can be explored in the future to improve the atlas generation. Summarizing,

the PRISMIN encoder generates coarse atlases containing emotions from EEG signals and significantly reduces the SEED dataset size with both GS and SR encodings even when applying a data augmentation strategy, thus satisfying the need for a smaller collection to train different types of deep learning models.

4.4. Empátheia classifier evaluation

This section presents ablation studies and a state-of-the-art comparison to demonstrate the effectiveness of the Empátheia system. Specifically, the former is discussed in Sec. 4.4.1, which examines various aspects of the implemented architectures. The literature comparison is presented in Sec. 4.4.2.

4.4.1. Ablation study

Extensive experiments were performed to fully evaluate the Empátheia classifiers as they are being applied to a novel input, i.e. the PRISMIN generated atlases. Specifically, ablation studies have been conducted on both GS and SR encodings, i.e. datasets \mathcal{D}_{GS} and \mathcal{D}_{SR} , with different kernel sizes ($K \in \{3, 5, 16\}$), learning rates ($LR \in \{1e-3, 2e-3, 5e-4\}$), and, for recurrent models, hidden units number ($U \in \{30, 50, 70\}$). Regarding the ViT model, the study has been conducted with different learning rates ($LR \in \{1e-4, 1e-5, 1e-6\}$) and number of heads ($\mathcal{H} \in \{8, 16, 32\}$) using a 16×16 patch size, and a 12-layers transformer encoder. The results obtained on the development set are reported in Table 2 for the Conv classifier, Tables 3–5 for the ConvLSTM classifier, Tables 6–8 for the ConvGRU classifier, and Table 9 for the ViT classifier. The best model in each table for the GS encoding is highlighted in green, while the best for the SR is in blue.

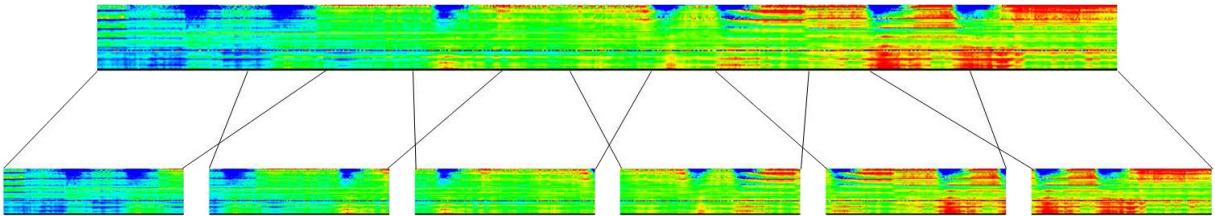


Fig. 5. Data augmentation strategy example.

Table 2. Conv classifier ablation.

\mathcal{D}	K	LR	Acc	Prec	Recall	F1
\mathcal{D}_{GS}	3	1e-3	69.7%	70.5%	69.7%	70.1%
\mathcal{D}_{SR}	3	1e-3	66.6%	65.9%	66.6%	66.3%
\mathcal{D}_{GS}	3	2e-3	58.1%	57.0%	58.0%	57.5%
\mathcal{D}_{SR}	3	2e-3	71.1%	71.5%	71.1%	71.5%
\mathcal{D}_{GS}	3	5e-4	63.5%	67.7%	63.5%	65.6%
\mathcal{D}_{SR}	3	5e-4	71.6%	71.3%	71.6%	71.5%
\mathcal{D}_{GS}	5	1e-3	66.8%	66.9%	66.8%	66.9%
\mathcal{D}_{SR}	5	1e-3	70.6%	72.6%	70.7%	71.6%
\mathcal{D}_{GS}	5	2e-3	57.1%	71.8%	57.1%	63.6%
\mathcal{D}_{SR}	5	2e-3	62.2%	72.4%	62.8%	67.3%
\mathcal{D}_{GS}	5	5e-4	68.9%	69.6%	68.9%	69.3%
\mathcal{D}_{SR}	5	5e-4	71.1%	71.5%	71.1%	71.3%
\mathcal{D}_{GS}	16	1e-3	63.4%	64.7%	63.4%	64.1%
\mathcal{D}_{SR}	16	1e-3	71.7%	72.8%	71.7%	71.3%
\mathcal{D}_{GS}	16	2e-3	54.8%	60.4%	54.8%	57.4%
\mathcal{D}_{SR}	16	2e-3	64.6%	65.7%	64.6%	65.1%
\mathcal{D}_{GS}	16	5e-4	65.1%	65.9%	65.1%	65.5%
\mathcal{D}_{SR}	16	5e-4	67.1%	69.5%	67.1%	68.3%

The choice of using a convolutional approach for extracting the features from the atlas comes from the benchmark presented in Avola *et al.*⁵⁷ In such benchmark, raw EEG signals have been tested with vanilla CNN, LSTM, and GRU, highlighting that CNNs achieve the best accuracy with such type of data. As observed in the various tables, all

Table 3. ConvLSTM classifier ablation, $LR = 1e - 3$.

\mathcal{D}	K	U	Acc	Prec	Recall	F1
\mathcal{D}_{GS}	3	30	77.3%	77.5%	77.3%	77.4%
\mathcal{D}_{SR}	3	30	73.7%	73.4%	73.8%	73.6%
\mathcal{D}_{GS}	3	50	79.1%	79.7%	79.1%	79.4%
\mathcal{D}_{SR}	3	50	75.4%	75.5%	75.1%	75.3%
\mathcal{D}_{GS}	3	70	80.1%	80.1%	80.1%	80.1%
\mathcal{D}_{SR}	3	70	76.4%	75.8%	76.1%	76.0%
\mathcal{D}_{GS}	5	30	77.7%	78.5%	77.8%	78.2%
\mathcal{D}_{SR}	5	30	74.5%	74.3%	73.9%	74.2%
\mathcal{D}_{GS}	5	50	77.5%	78.0%	77.6%	77.8%
\mathcal{D}_{SR}	5	50	74.5%	73.8%	73.7%	73.8%
\mathcal{D}_{GS}	5	70	74.8%	74.2%	73.9%	74.0%
\mathcal{D}_{SR}	5	70	71.6%	70.2%	70.2%	70.2%
\mathcal{D}_{GS}	16	30	80.0%	80.6%	80.0%	80.3%
\mathcal{D}_{SR}	16	30	76.4%	76.3%	76.0%	76.2%
\mathcal{D}_{GS}	16	50	79.7%	79.5%	79.8%	79.6%
\mathcal{D}_{SR}	16	50	76.4%	75.3%	75.8%	75.5%
\mathcal{D}_{GS}	16	70	82.1%	82.0%	82.1%	82.1%
\mathcal{D}_{SR}	16	70	77.6%	77.8%	77.7%	77.8%

Table 4. ConvLSTM classifier ablation, $LR = 2e-3$.

\mathcal{D}	K	U	Acc	Prec	Recall	F1
\mathcal{D}_{GS}	3	30	75.9%	76.4%	76.0%	76.2%
\mathcal{D}_{SR}	3	30	62.1%	62.2%	62.1%	62.2%
\mathcal{D}_{GS}	3	50	78.6%	78.6%	78.6%	78.6%
\mathcal{D}_{SR}	3	50	69.5%	69.5%	69.4%	69.5%
\mathcal{D}_{GS}	3	70	77.0%	77.2%	77.0%	77.1%
\mathcal{D}_{SR}	3	70	68.1%	68.3%	68.0%	68.2%
\mathcal{D}_{GS}	5	30	75.6%	75.5%	75.7%	75.6%
\mathcal{D}_{SR}	5	30	66.8%	66.8%	66.9%	66.8%
\mathcal{D}_{GS}	5	50	75.0%	75.0%	75.0%	75.0%
\mathcal{D}_{SR}	5	50	66.3%	66.3%	66.3%	66.3%
\mathcal{D}_{GS}	5	70	77.8%	78.2%	77.9%	78.0%
\mathcal{D}_{SR}	5	70	68.8%	69.2%	68.8%	69.0%
\mathcal{D}_{GS}	16	30	74.0%	74.3%	74.1%	74.2%
\mathcal{D}_{SR}	16	30	65.4%	65.7%	65.5%	65.6%
\mathcal{D}_{GS}	16	50	77.2%	77.8%	77.2%	77.5%
\mathcal{D}_{SR}	16	50	68.3%	68.8%	68.2%	68.5%
\mathcal{D}_{GS}	16	70	76.8%	76.6%	76.8%	76.7%
\mathcal{D}_{SR}	16	70	73.0%	73.2%	73.0%	73.1%

Empátheia classifiers achieve significant performance on the emotion recognition task using the generated atlases, reaching up to 83.5% accuracy. Regarding convolutional-based classifiers, the Conv model falls slightly behind the ConvLSTM and ConvGRU architectures, reporting accuracy gaps of 13.8% and 8.6% on the \mathcal{D}_{GS} and \mathcal{D}_{SR} datasets, respectively.

Table 5. ConvLSTM classifier ablation, $LR = 5e-4$.

\mathcal{D}	K	U	Acc	Prec	Recall	F1
\mathcal{D}_{GS}	3	30	80.6%	80.7%	80.6%	80.7%
\mathcal{D}_{SR}	3	30	73.9%	74.2%	74.0%	74.1%
\mathcal{D}_{GS}	3	50	80.6%	80.5%	80.6%	80.6%
\mathcal{D}_{SR}	3	50	76.3%	74.4%	74.4%	74.4%
\mathcal{D}_{GS}	3	70	80.5%	80.9%	80.7%	80.8%
\mathcal{D}_{SR}	3	70	75.5%	74.8%	74.5%	74.6%
\mathcal{D}_{GS}	5	30	78.7%	78.4%	78.7%	78.6%
\mathcal{D}_{SR}	5	30	74.4%	72.5%	72.7%	72.6%
\mathcal{D}_{GS}	5	50	82.8%	83.0%	82.9%	82.9%
\mathcal{D}_{SR}	5	50	76.4%	76.7%	76.6%	76.6%
\mathcal{D}_{GS}	5	70	76.1%	76.2%	76.2%	76.2%
\mathcal{D}_{SR}	5	70	71.6%	70.4%	70.4%	70.4%
\mathcal{D}_{GS}	16	30	79.6%	80.0%	79.7%	79.8%
\mathcal{D}_{SR}	16	30	75.4%	73.9%	73.6%	73.7%
\mathcal{D}_{GS}	16	50	80.0%	80.7%	80.0%	80.3%
\mathcal{D}_{SR}	16	50	75.4%	74.6%	73.9%	74.2%
\mathcal{D}_{GS}	16	70	81.7%	81.8%	81.8%	81.8%
\mathcal{D}_{SR}	16	70	76.0%	76.1%	76.1%	76.1%

Table 6. ConvGRU classifier ablation, $LR = 1e-3$.

\mathcal{D}	K	U	Acc	Prec	Recall	F1
\mathcal{D}_{GS}	3	30	71.7%	75.3%	71.7%	73.5%
\mathcal{D}_{SR}	3	30	67.8%	68.1%	67.8%	68.0%
\mathcal{D}_{GS}	3	50	76.5%	77.7%	76.5%	77.1%
\mathcal{D}_{SR}	3	50	73.3%	73.6%	74.3%	73.4%
\mathcal{D}_{GS}	3	70	80.3%	80.7%	77.8%	79.2%
\mathcal{D}_{SR}	3	70	77.5%	76.5%	75.5%	75.4%
\mathcal{D}_{GS}	5	30	79.2%	79.5%	79.3%	79.4%
\mathcal{D}_{SR}	5	30	76.3%	75.3%	77.0%	75.6%
\mathcal{D}_{GS}	5	50	79.7%	79.7%	79.8%	79.8%
\mathcal{D}_{SR}	5	50	77.0%	75.5%	77.5%	76.0%
\mathcal{D}_{GS}	5	70	77.1%	78.0%	77.1%	77.6%
\mathcal{D}_{SR}	5	70	74.3%	73.9%	74.9%	73.9%
\mathcal{D}_{GS}	16	30	79.3%	79.6%	79.4%	79.5%
\mathcal{D}_{SR}	16	30	76.3%	75.4%	77.1%	75.5%
\mathcal{D}_{GS}	16	50	77.7%	77.8%	77.8%	77.8%
\mathcal{D}_{SR}	16	50	75.3%	73.7%	75.5%	74.1%
\mathcal{D}_{GS}	16	70	75.5%	76.6%	75.6%	76.1%
\mathcal{D}_{SR}	16	70	75.3%	75.9%	75.3%	75.6%

The results obtained indicate that recurrent models can better capture the time evolution of emotions within the generated atlases through their recurrent layers. However, the kernel size in convolutional layers seems to play a fundamental role in achieving higher performance. In fact, when examining ConvLSTM and ConvGRU models, they tend to consistently achieve higher performance with wider

Table 7. ConvGRU classifier ablation, $LR = 2e-3$.

\mathcal{D}	K	U	Acc	Prec	Recall	F1
\mathcal{D}_{GS}	3	30	76.7%	76.5%	76.7%	76.6%
\mathcal{D}_{SR}	3	30	71.9%	71.4%	72.0%	71.7%
\mathcal{D}_{GS}	3	50	73.9%	74.6%	74.0%	74.3%
\mathcal{D}_{SR}	3	50	67.1%	67.8%	67.0%	67.4%
\mathcal{D}_{GS}	3	70	66.1%	69.7%	66.1%	67.9%
\mathcal{D}_{SR}	3	70	59.8%	63.4%	59.8%	61.6%
\mathcal{D}_{GS}	5	30	73.0%	72.8%	73.0%	72.9%
\mathcal{D}_{SR}	5	30	66.2%	66.2%	66.1%	66.1%
\mathcal{D}_{GS}	5	50	73.8%	73.7%	73.9%	73.8%
\mathcal{D}_{SR}	5	50	67.1%	67.0%	66.9%	66.9%
\mathcal{D}_{GS}	5	70	73.0%	74.9%	73.0%	73.9%
\mathcal{D}_{SR}	5	70	66.2%	68.1%	66.1%	67.0%
\mathcal{D}_{GS}	16	30	81.9%	82.5%	81.9%	82.2%
\mathcal{D}_{SR}	16	30	74.3%	75.0%	74.2%	74.5%
\mathcal{D}_{GS}	16	50	82.8%	83.0%	82.9%	82.9%
\mathcal{D}_{SR}	16	50	75.1%	75.5%	75.1%	75.2%
\mathcal{D}_{GS}	16	70	81.7%	82.1%	81.8%	82.0%
\mathcal{D}_{SR}	16	70	71.6%	72.9%	71.6%	72.2%

Table 8. ConvGRU classifier ablation, $LR = 5e-4$.

\mathcal{D}	K	U	Acc	Prec	Recall	F1
\mathcal{D}_{GS}	3	30	82.2%	82.5%	82.2%	82.4%
\mathcal{D}_{SR}	3	30	73.4%	73.7%	73.4%	73.6%
\mathcal{D}_{GS}	3	50	79.6%	80.9%	79.7%	80.3%
\mathcal{D}_{SR}	3	50	72.4%	79.4%	79.6%	79.5%
\mathcal{D}_{GS}	3	70	83.5%	83.9%	83.6%	83.7%
\mathcal{D}_{SR}	3	70	80.3%	78.1%	76.6%	77.3%
\mathcal{D}_{GS}	5	30	77.8%	78.0%	76.8%	77.4%
\mathcal{D}_{SR}	5	30	69.6%	72.6%	70.4%	71.5%
\mathcal{D}_{GS}	5	50	71.8%	73.5%	71.9%	72.7%
\mathcal{D}_{SR}	5	50	65.1%	68.4%	65.9%	67.2%
\mathcal{D}_{GS}	5	70	77.0%	77.3%	77.0%	77.2%
\mathcal{D}_{SR}	5	70	69.6%	72.0%	70.6%	71.3%
\mathcal{D}_{GS}	16	30	75.5%	75.3%	75.6%	75.4%
\mathcal{D}_{SR}	16	30	68.7%	70.1%	69.3%	69.7%
\mathcal{D}_{GS}	16	50	72.2%	74.2%	72.3%	73.2%
\mathcal{D}_{SR}	16	50	65.2%	69.1%	66.3%	67.6%
\mathcal{D}_{GS}	16	70	81.0%	81.5%	81.0%	81.2%
\mathcal{D}_{SR}	16	70	69.8%	75.1%	69.8%	72.4%

kernels (i.e. $K = 16$). This suggests that temporal information is still captured by their convolutional receptive fields. Other relevant hyper-parameters that strongly impact the final performance include the learning rate and, for recurrent models, the number of hidden units. Regarding the learning rate, it directly affects model convergence and shows improved metrics when smaller LR values are

Table 9. ViT classifier ablation.

\mathcal{D}	\mathcal{H}	LR	Acc	Prec	Recall	F1
\mathcal{D}_{GS}	8	1e-4	56.4%	51.5%	53.8%	52.6%
\mathcal{D}_{SR}	8	1e-4	57.1%	52.3%	54.3%	53.3%
\mathcal{D}_{GS}	8	1e-5	55.7%	52.4%	52.12%	52.78%
\mathcal{D}_{SR}	8	1e-5	60.0%	56.7%	57.5%	57.1%
\mathcal{D}_{GS}	8	1e-6	56.1%	53.3%	53.5%	53.4%
\mathcal{D}_{SR}	8	1e-6	53.7%	52.4%	52.0%	52.2%
\mathcal{D}_{GS}	16	1e-4	58.2%	55.2%	55.4%	55.3%
\mathcal{D}_{SR}	16	1e-4	59.7%	57.7%	58.4%	58.1%
\mathcal{D}_{GS}	16	1e-5	56.2%	53.8%	54.4%	54.1%
\mathcal{D}_{SR}	16	1e-5	58.8%	55.7%	56.6%	56.1%
\mathcal{D}_{GS}	16	1e-6	54.7%	52.8%	53.2%	53.0%
\mathcal{D}_{SR}	16	1e-6	52.7%	50.5%	51.3%	50.9%
\mathcal{D}_{GS}	32	1e-4	57.3%	57.7%	56.9%	57.3%
\mathcal{D}_{SR}	32	1e-4	60.0%	59.0%	59.6%	59.4%
\mathcal{D}_{GS}	32	1e-5	55.0%	56.3%	54.6%	55.4%
\mathcal{D}_{SR}	32	1e-5	55.5%	56.0%	55.3%	55.6%
\mathcal{D}_{GS}	32	1e-6	60.0%	59.7%	59.6%	59.6%
\mathcal{D}_{SR}	32	1e-6	53.9%	53.0%	53.6%	53.3%

associated with larger kernels and vice versa. This result indicates that moving more slowly along the error surface during training, i.e. with smaller LR values (e.g. $5e-4$), allows the model to thoroughly analyze the broader receptive fields along with the temporal information. On the other hand, moving faster along the error surface, i.e. with larger LR values (e.g. $2e-3$), avoids fixating on details captured by the smaller kernels that might be associated with noise since the atlases provide a coarse representation of the EEG signals. With respect to the hidden units size of recurrent architectures, i.e. ConvLSTM and ConvGRU, a U value of either 50 or 70 appears to be the preferred size to capture all emotion-related details within an atlas, especially when combined with lower LR values, as can be observed, among others, in Table 8. Regarding ViT, it falls behind all Conv-based architectures, reporting an accuracy downgrade of 23.5% and 20.3% compared to the ConvGRU on the \mathcal{D}_{GS} and \mathcal{D}_{SR} datasets, respectively. However, the number of heads seems to play a significant role in achieving higher performance. In fact, as can be seen in Table 9, ViT tends to achieve higher performance with a higher number of attention heads (i.e. $\mathcal{H} = 32$). This suggests that a more complex model is likely to perform better on the proposed encodings. The last aspect affecting model performance is related to the type of

encoding used to create the atlases. Specifically, the Conv model achieves higher performance when using atlases from \mathcal{D}_{SR} as input. On the other hand, recurrent classifiers, namely, ConvLSTM and ConvGRU, yield better metrics when evaluated on \mathcal{D}_{GS} . This suggests that the temporal evolution of emotions, which is better captured by architectures with recurrent components, is more effectively encoded through GS mapping. This also implies that employing alternative encodings could potentially yield significantly different results based on the underlying architecture.

Independently of the underlying hyper-parameters configuration, all Empátheia classifiers demonstrate stable performance across all classification metrics. In fact, accuracy, precision, recall, and F1 show aligned values suggesting that the models provide balanced performance without bias towards any particular class, avoiding false positives and effectively identifying positive instances for each emotion. This indicates that the learned weights are robust and can correctly abstract the emotion represented in the PRISMIN-extracted atlas. This outcome can also be observed in Figs. 6 and 7, which depict, respectively, accuracy curves during training using the best hyper-parameters for each model and the confusion matrices computed on the test set. As shown in Fig. 7, recurrent models maintain high

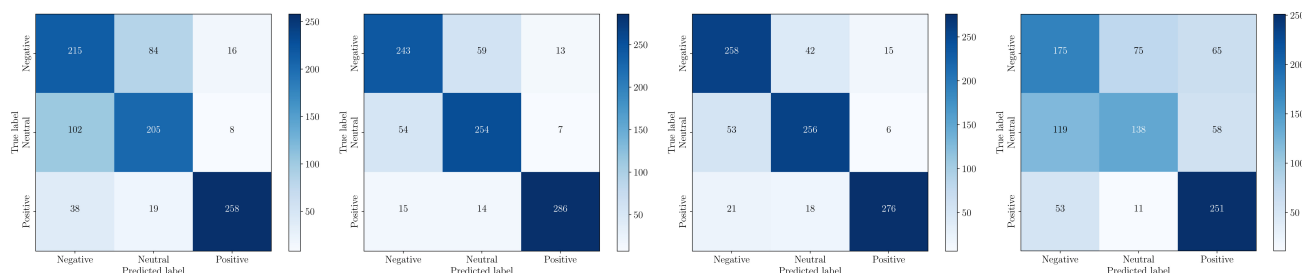


Fig. 6. Test set confusion matrices, from left to right: results of Conv, ConvLSTM, ConvGRU, and ViT models.

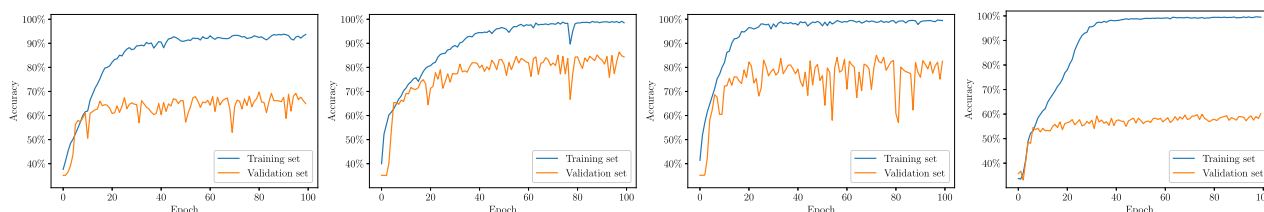


Fig. 7. Training set accuracies, from left to right: results of Conv, ConvLSTM, ConvGRU, and ViT models.

performance on the development set throughout their training. Conversely, ViT and Conv classifiers suffer from more prominent overfitting and result in reduced metrics compared to the other models. Regarding CNN, this behavior becomes more apparent in the confusion matrices related to the test set, where the Conv classifier makes more mistakes. Concerning ViT, as pointed out in Khan *et al.*,⁵⁸ transformer architectures lack inherent encoding of inductive biases (prior knowledge) for handling visual data. They typically require large amounts of training data to discern the underlying modality-specific rules. This increased complexity is a result of the larger number of parameters, as shown in Table 10. For this reason, the ViT model presents the highest overfitting value (around 40%), requiring much more data to generalize better. In fact, unlike CNN-based models equipped with built-in translation invariance, weight sharing, and partial scale invariance, transformer networks are required to deduce these image-specific concepts from the provided training examples autonomously. Based on the experimental results, the ViT architecture is proved to be unsuitable for the scope of the work. In fact, compared to other models, ViT not only exhibits the lowest classification accuracy but also demands heavy computational resources for parameter and weight management, as reported in Table 10. Interestingly, all models seem to mismatch samples that are mostly associated with negative and neutral emotions. This suggests that these categories share common patterns in the generated atlases, indicating that further exploration on the atlas generation by the PRISMIN encoder might improve the final classification. Finally, Table 10 compares the best

configurations among the reported ablation studies to underline the effectiveness of the devised solutions in classifying the GS and SR encodings. As can be observed, recurrent models achieve the highest performance, with ConvGRU being the best model on both \mathcal{D}_{SR} and \mathcal{D}_{GS} datasets. This demonstrates that temporal information is preserved in the coarse representation of EEG signals transformed into atlases, indicating that architectures with recurrent elements can exhibit varying performance depending on internal design but should be the preferred choice, especially when applied to the proposed input representation.

4.4.2. State-of-the-art comparison

To conclude the evaluation of the Empátheia classifier, a state-of-the-art comparison was conducted using the SEED dataset. The results are presented in Table 11. As can be observed, the Empátheia system achieves comparable performance with many existing works in the literature. This result is intriguing, considering that the generated atlases represent a coarse transformation of the EEG signals used by other approaches. This suggests potential room for improvement in the presented PRISMIN encoder. Furthermore, existing solutions, even the highest-performing ones, employ advanced yet complex models to handle the fine-grained details of EEG signals. However, this complexity comes at the cost of increased computational demands and the need for larger datasets. In contrast, as demonstrated in Table 1, the Empátheia classifiers achieve the reported performance with a dataset of lower quality that requires less disk space to be stored, enabling

Table 10. Best ablation configurations comparison.

Model	\mathcal{D}	Acc	Prec	Recall	F1	FLOPS*	Params
Conv	\mathcal{D}_{GS}	69.7%	70.5%	69.7%	70.1%	0.1G	28.6M
Conv	\mathcal{D}_{SR}	71.7%	72.8%	71.7%	71.3%		
ViT	\mathcal{D}_{GS}	75.5%	75.4%	75.5%	75.4%	40.16G	86M
ViT	\mathcal{D}_{SR}	72.6%	72.1%	72.2%	72.2%		
ConvLSTM	\mathcal{D}_{GS}	82.8%	83.0%	82.9%	82.9%	0.037G	0.13M
ConvLSTM	\mathcal{D}_{SR}	77.6%	77.8%	77.7%	77.8%		
ConvGRU	\mathcal{D}_{GS}	83.5%	83.9%	83.6%	83.7%	0.035G	0.14M
ConvGRU	\mathcal{D}_{SR}	80.3%	78.1%	76.6%	77.3%		

Note: *Number of floating point operations per input.

Table 11. State-of-the-art performance comparison.

Model	Acc	Prec	Recall	F1
MFBPST-3D ³⁹	96.79%	—	—	—
DNN-SAE ³⁶	96.77%	—	—	—
TANN ⁵⁹	93.34%	—	—	—
SVM ²³	86.65%	—	—	—
DBN ²³	86.08%	—	—	—
ConvGRU	83.50%	83.90%	83.60%	83.70%
DTCW-SRU ⁵⁶	83.13%	82.24%	81.53%	81.24%
ConvLSTM	82.80%	83.00%	82.90%	82.90%
LRM ²³	82.70%	—	—	—
KNN ²³	72.60%	—	—	—
Conv	71.70%	72.80%	71.70%	71.30%
ViT	60.00%	59.70%	59.60%	59.60%

Table 12. State-of-the-art computational comparison.

Model	Params	Time*	GPU	FLOPS
MFBPST-3D ³⁹	9M	33s	RTX3090	—
DTCW-SRU ⁵⁶	2.45M	27s	RTX3090	—
ConvGRU	0.3M	0.66s	Quadro RTX6000	0.6G

Note: *Training time required to analyze a single trial.

the use of lightweight models. For instance, Table 12 compares the best-performing Empátheia classifier, ConvGRU, with existing architectures from a computational point of view. Not only is the proposed model noticeably smaller than existing solutions, but it also processes individual trials significantly faster, i.e. about 45 times faster, using similar hardware. This emphasizes the rationale behind the Empátheia system and highlights the advantage of its PRISMIN encoder, which reduces the dataset size and, consequently, improves computational efficiency in terms of model size and training speed, thereby demonstrating the effectiveness of the proposed approach.

5. Conclusion

This paper presents the Empátheia system, which performs emotion classification from EEG signals using a reduced amount of data. The proposed approach consists of two main components: the PRISMIN encoder and the Empátheia classifier. The


PRISMIN encoder generates coarse atlases — 2D images encoded using GS or SR mapping — to represent the captured emotion within EEG signals. This representation significantly reduces the input dataset, enabling the implementation of lightweight models and faster training times. The second component, the Empátheia classifier, utilizes DNNs tailored to capture spatio-temporal characteristics present in the atlases to perform emotion classification. Multiple and different tests on public reference datasets, i.e. SEED, have been performed to define the most accurate network that could classify emotions. The PRISMIN encoder reduced the dataset size by up to 10.3 times. The Empátheia classifiers achieved competitive performance in line with several existing literature works. However, due to the coarse nature of the atlas representation, they fell short of the best-performing approaches. Nevertheless, these results suggest ample room for improvement in both the PRISMIN encoder and Empátheia classifier. Furthermore, ablation studies on various Empátheia classifier hyper-parameters revealed stable performances across all experiments. The mixed models exhibited higher metrics due to their internal configurations, demonstrating the atlas ability to accurately represent emotion from EEG signals. The performed experiments showed that, among the tested models, the ConvGRU is the one achieving the best results.


As a potential avenue for future research, exploring additional encoding strategies is warranted, given their varying performance depending on the underlying architecture. This indicates that new strategies for the PRISMIN encoder could yield improved results. Additionally, implementing more advanced models as Empátheia classifiers could potentially compensate for the coarse representation derived from the atlases, for example, integrating the capability of recurrent architectures to capture temporal information with the attention mechanism.^{60,61} Also, some recent techniques, e.g. Neural Dynamic Classification (NDC) algorithms,⁶² Dynamic Ensemble Learning (DEL) approaches,⁶³ Finite Element Machine (FEM),⁶⁴ DA strategies,⁶⁵ Functional Connectivity (FC),⁶⁶ and self-supervised learning⁶⁷ could provide a guideline about possible strategies to apply to compressed signals.


Acknowledgments


This work was supported by “Smart unmannEd AeRial vehiCles for Human likE monitoRing (SEARCHER)” project of the Italian Ministry of Defence within the PNRM 2020 Program (Grant No. PNRM a2020.231); and “A Brain–Computer Interface (BCI)-based System for Transferring Human Emotions inside Unmanned Aerial Vehicles (UAVs)” Sapienza University Research Projects (Grant No. RM1221816C1CF63B); and Departmental Strategic Plan (DSP) of the University of Udine — Interdepartmental Project on Artificial Intelligence (2020–25); and “A proactive counter-UAV system to protect army tanks and patrols in urban areas (PROACTIVE COUNTER-UAV)” project of the Italian Ministry of Defence (No. 2066/16.12.2019); and the Made in Italy Circular and Sustainable (MICS) Extended Partnership and received funding from Next-Generation EU (Italian PNRR M4 C2, Invest 1.3 D.D. 1551.11-10-2022, PE00000004). CUP MICS B53C22004130001; and PNRR project “Humanities and Heritage Italian Open Science Cloud” - H2IOSC, CUP B63C22000730005; and “EYE-FI.AI: going bEYond computEr vision paradigm using wi-FI signals in AI systems” project of the Italian Ministry of Universities and Research (MUR) within the PRIN 2022 Program (Grant Number: 2022AL45R2) (CUP: B53D23012950001).

ORCID


Daniilo Avola  <https://orcid.org/0000-0001-9437-6217>


Luigi Cinque  <https://orcid.org/0000-0001-9149-2175>


Angelo Di Mambro  <https://orcid.org/0000-0001-9834-3519>

Alessio Fagioli  <https://orcid.org/0000-0002-8111-9120>

Marco Raoul Marini  <https://orcid.org/0000-0002-2540-2570>

Daniele Pannone  <https://orcid.org/0000-0001-6446-6473>

Bruno Fanini  <https://orcid.org/0000-0003-4058-877X>

Gian Luca Foresti  <https://orcid.org/0000-0002-8425-6892>

References

1. N. Palomero-Gallagher and K. Amunts, A short review on emotion processing: A lateralized network of neuronal networks, *Brain Struct. Funct.* **227**(2) (2022) 673–684.
2. L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez and S. D. Pollak, Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements, *Psychol. Sci. Public Interest* **20**(1) (2019) 1–68.
3. S. Passardi, P. Peyk, M. Rufer, T. S. Wingenbach and M. C. Pfaltz, Facial mimicry, facial emotion recognition and alexithymia in post-traumatic stress disorder, *Behav. Res. Ther.* **122** (2019) 1–8.
4. A. Semwal and N. D. Londhe, Computer aided pain detection and intensity estimation using compact CNN based fusion network, *Appl. Soft Comput.* **112** (2021) 1–11.
5. R. Yuvaraj, M. Murugappan, U. R. Acharya, H. Adeli, N. M. Ibrahim and E. Mesquita, Brain functional connectivity patterns for emotional state classification in Parkinson’s disease patients without dementia, *Behav. Brain Res.* **298**(Pt B) (2015) 248–260.
6. D. Avola, L. Cinque, M. De Marsico, A. Fagioli and G. L. Foresti, LieToMe: Preliminary study on hand gestures for deception detection via fisher-LSTM, *Pattern Recognit. Lett.* **138** (2020) 455–461.
7. D. Avola, M. Cascio, L. Cinque, A. Fagioli and G. L. Foresti, LieToMe: An ensemble approach for deception detection from facial cues, *Int. J. Neural Syst.* **31**(02) (2021) 2050068.
8. M. A. Vicente-Querol, A. Fernández-Caballero, J. P. Molina, L. M. González-Gualda, P. Fernández-Sotos and A. S. García, Facial affect recognition in immersive virtual reality: Where is the participant looking? *Int. J. Neural Syst.* **32**(10) (2022) 2250029.
9. D. Avola, L. Cinque, A. Fagioli, G. L. Foresti and C. Massaroni, Deep temporal analysis for non-acted body affect recognition, *IEEE Trans. Affect. Comput.* **13**(3) (2020) 1366–1377.
10. D. Avola, M. Cascio, L. Cinque, A. Fagioli and G. L. Foresti, Affective action and interaction recognition by multi-view representation learning from hand-crafted low-level skeleton features, *Int. J. Neural Syst.* **32**(10) (2022) 2250040.
11. J. De Lope and M. Graña, A hybrid Time-Distributed deep neural architecture for speech emotion recognition, *Int. J. Neural Syst.* **32**(6) (2022) 2250024.
12. A. Nandi, F. Khafa, L. Subirats and S. Fort, Reward-penalty weighted ensemble for emotion state classification from multi-modal data streams, *Int. J. Neural Syst.* **32**(12) (2022) 2250049.
13. A. Burns and H. Adeli, Wearable technology for patients with brain and spinal cord injuries, *Rev. Neurosci.* **28**(8) (2017) 913–920.

14. A. Ortiz-Rosario and H. Adeli, Brain-computer interface technologies: From signal to action, *Rev. Neurosci.* **24**(5) (2013) 537–552.
15. X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar and P. Marttinen, EEG based emotion recognition: A tutorial and review, *ACM Comput. Surv.* **55**(4) (2022) 1–57.
16. A. Olamat, P. Ozel and S. Atasever, Deep learning methods for multi-channel EEG-based emotion recognition, *Int. J. Neural Syst.* **32**(05) (2022) 2250021.
17. D. W. Prabowo, H. A. Nugroho, N. A. Setiawan and J. Debayle, A systematic literature review of emotion recognition using EEG signals, *Cogn. Syst. Res.* **82** (2023) 101152.
18. T. Harmony, T. Fernández, J. Silva, J. Bernal, L. Díaz-Comas, A. Reyes, E. Marosi, M. Rodríguez and M. Rodríguez, EEG delta activity: An indicator of attention to internal processing during performance of mental tasks, *Int. J. Psychophysiol.* **24**(1–2) (1996) 161–171.
19. W. Klimesch, EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis, *Brain Res. Rev.* **29**(2–3) (1999) 169–195.
20. W. Klimesch, EEG-alpha rhythms and memory processes, *Int. J. Psychophysiol.* **26**(1–3) (1997) 319–340.
21. A. K. Engel and P. Fries, Beta-band oscillations—signalling the status quo? *Curr. Opin. Neurobiol.* **20**(2) (2010) 156–165.
22. W. H. Miltner, C. Braun, M. Arnold, H. Witte and E. Taub, Coherence of gamma-band EEG activity as a basis for associative learning, *Nature* **397**(6718) (1999) 434–436.
23. W.-L. Zheng and B.-L. Lu, Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks, *IEEE Trans. Auton. Ment. Dev.* **7**(3) (2015) 162–175.
24. X. Gu, Z. Cao, A. Jolfaei, P. Xu, D. Wu, T.-P. Jung and C.-T. Lin, EEG-based brain-computer interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications, *IEEE/ACM Trans. Comput. Biol. Bioinform.* **18**(5) (2021) 1645–1666.
25. J. Singh, F. Ali, R. Gill, B. Shah and D. Kwak, A survey of EEG and machine learning-based methods for neural rehabilitation, *IEEE Access* **11** (2023) 114155–114171.
26. A. Othmani, B. Brahem, Y. Haddou and Mustaqeem, Machine-learning-based approaches for post-traumatic stress disorder diagnosis using video and EEG sensors: A review, *IEEE Sens. J.* **23**(20) (2023) 24135–24151.
27. A. Bablani, D. R. Edla, D. Tripathi and R. Cheruku, Survey on brain-computer interface: An emerging computational intelligence paradigm, *ACM Comput. Surv.* **52**(1) (2019) 1–32.
28. B. Fanini and L. Cinque, Encoding, exchange and manipulation of captured immersive VR sessions for learning environments: The PRISMIN framework, *Appl. Sci.* **10**(6) (2020) 1–16.
29. R.-N. Duan, J.-Y. Zhu and B.-L. Lu, Differential entropy feature for EEG-based emotion classification, in *Proc. Int. IEEE/EMBS Conf. Neural Engineering (NER)* (IEEE, 2013), pp. 81–84.
30. A. Khosla, P. Khandnor and T. Chand, A comparative analysis of signal processing and classification methods for different applications based on EEG signals, *Bio-cybern. Biomed. Eng.* **40**(2) (2020) 649–690.
31. S. Lemm, B. Blankertz, G. Curio and K.-R. Müller, Spatio-spectral filters for improving the classification of single trial EEG, *IEEE Trans. Biomed. Eng.* **52**(9) (2005) 1541–1548.
32. R. Jenke, A. Peer and M. Buss, Feature extraction and selection for emotion recognition from EEG, *IEEE Trans. Affect. Comput.* **5**(3) (2014) 327–339.
33. Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann and J.-H. Chen, EEG-based emotion recognition in music listening, *IEEE Trans. Biomed. Eng.* **57**(7) (2010) 1798–1806.
34. M. Akin, Comparison of wavelet transform and FFT methods in the analysis of EEG signals, *J. Med. Syst.* **26** (2002) 241–247.
35. E. H. Houssein, A. Hammad and A. A. Ali, Human emotion recognition from EEG-based brain-computer interface using machine learning: A comprehensive review, *Neural. Comput. Appl.* **34**(15) (2022) 12527–12557.
36. J. Liu, G. Wu, Y. Luo, S. Qiu, S. Yang, W. Li and Y. Bi, EEG-based emotion classification using a deep neural network and sparse autoencoder, *Front. Syst. Neurosci.* **14** (2020) 1–43.
37. S. Liu, Z. Wang, Y. An, J. Zhao, Y. Zhao and Y.-D. Zhang, EEG emotion recognition based on the attention mechanism and pre-trained convolution capsule network, *Knowl.-Based Syst.* **265** (2023) 110372.
38. D. Li, L. Xie, Z. Wang and H. Yang, Brain emotion perception inspired EEG emotion recognition with deep reinforcement learning, *IEEE Trans. Neural Netw. Learn. Syst.* (2023) 1–14, <https://ieeexplore.ieee.org/document/10113206>.
39. M. Miao, L. Zheng, B. Xu, Z. Yang and W. Hu, A multiple frequency bands parallel spatial-temporal 3D deep residual learning framework for EEG-based emotion recognition, *Biomed. Signal Process. Control* **79** (2023) 104141.
40. J. Hu, C. Wang, Q. Jia, Q. Bu, R. Sutcliffe and J. Feng, ScalingNet: Extracting features from raw EEG data for emotion recognition, *Neurocomputing* **463** (2021) 177–184.
41. Y. Li, L. Wang, W. Zheng, Y. Zong, L. Qi, Z. Cui, T. Zhang and T. Song, A novel bi-hemispheric discrepancy model for EEG emotion recognition, *IEEE Trans. Cogn. Develop. Syst.* **13**(2) (2020) 354–367.

42. T. Zhang, W. Zheng, Z. Cui, Y. Zong and Y. Li, Spatial-temporal recurrent neural network for emotion recognition, *IEEE Trans. Cybern.* **49**(3) (2018) 839–847.
43. W. Guo, G. Xu and Y. Wang, Multi-source domain adaptation with spatio-temporal feature extractor for EEG emotion recognition, *Biomed. Signal Process. Control* **84** (2023) 104998.
44. Y. Yang, Q. Wu, M. Qiu, Y. Wang and X. Chen, Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network, in *Proc. Int. Joint Conf. Neural Networks (IJCNN)* (IEEE, 2018), pp. 1–7.
45. X. Du, C. Ma, G. Zhang, J. Li, Y.-K. Lai, G. Zhao, X. Deng, Y.-J. Liu and H. Wang, An efficient LSTM network for emotion recognition from multichannel EEG signals, *IEEE Trans. Affect. Comput.* **13**(3) (2020) 1528–1540.
46. S. Liu, Y. Zhao, Y. An, J. Zhao and S.-H. Wang and J. Yan, GLFANet: A global to local feature aggregation network for EEG emotion recognition, *Biomed. Signal Process. Control* **85** (2023) 104799.
47. P. Zhong, D. Wang and C. Miao, EEG-based emotion recognition using regularized graph neural networks, *IEEE Trans. Affect. Comput.* **13**(3) (2020) 1290–1301.
48. T. Song, W. Zheng, P. Song and Z. Cui, EEG emotion recognition using dynamical graph convolutional neural networks, *IEEE Trans. Affect. Comput.* **11**(3) (2018) 532–541.
49. Y. Zhou, F. Li, Y. Li, Y. Ji, G. Shi, W. Zheng, L. Zhang, Y. Chen and R. Cheng, Progressive graph convolution network for EEG emotion recognition, *Neurocomputing* **544** (2023) 126262.
50. Y. Yin, X. Zheng, B. Hu, Y. Zhang and X. Cui, EEG emotion recognition using fusion model of graph convolutional neural networks and LSTN, *Appl. Soft Comput.* **100** (2021) 106954.
51. M. Limper, Y. Jung, J. Behr, T. Sturm, T. Franke, K. Schwenk and A. Kuijper, Fast, progressive loading of binary-encoded declarative-3D web content, *IEEE Comput. Graph. Appl.* **33**(5) (2013) 26–36.
52. B. Fanini and L. Cinque, Encoding immersive sessions for online, interactive VR analytics, *Virtual Real.* **24** (3) (2020) 423–438.
53. A. Graves and J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Netw.* **18**(5–6) (2005) 602–610.
54. A. Dosovitskiy et al., An image is worth 16×16 words: Transformers for image recognition at scale, arXiv:2010.11929.
55. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, Vol. 30, eds. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Curran Associates, 2017), pp. 6000–6010.
56. C. Wei, L.-L. Chen, Z.-Z. Song, X.-G. Lou and D.-D. Li, EEG-based emotion recognition using simple recurrent units network and ensemble learning, *Biomed. Signal Process. Control* **58** (2020) 101756.
57. D. Avola, M. Cascio, L. Cinque, A. Fagioli, G. L. Foresti, M. R. Marini and D. Pannone, Analyzing EEG data with machine and deep learning: A benchmark, in *Proc. Int. Conf. Neural Image Analysis and Processing (ICIAP)* (Springer, Cham, 2022), pp. 335–345.
58. S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan and M. Shah, Transformers in vision: A survey, *ACM Comput. Surv.* **54**(10s) (2022) 1–41.
59. Y. Li, B. Fu, F. Li, G. Shi and W. Zheng, A novel transferability attention neural network model for EEG emotion recognition, *Neurocomputing* **447** (2021) 92–101.
60. A. R. Javed, S. Ur Rehman, M. U. Khan, M. Alazab and T. Reddy, CANintelliIDS: Detecting in-vehicle intrusion attacks on a controller area network using CNN and attention-based GRU, *IEEE Trans. Netw. Sci. Eng.* **8**(2) (2021) 1456–1466.
61. X. Yin, Z. Liu, D. Liu and X. Ren, A novel CNN-based bi-LSTM parallel model with attention mechanism for human activity recognition with noisy data, *Sci. Rep.* **12**(1) (2022) 7878.
62. M. H. Rafiei and H. Adeli, A new neural dynamic classification algorithm, *IEEE Trans. Neural Netw. Learn. Syst.* **28**(12) (2017) 3074–3083.
63. K. M. R. Alam, N. Siddique and H. Adeli, A dynamic ensemble learning algorithm for neural networks, *Neural Comput. Appl.* **32**(12) (2020) 8675–8690.
64. D. R. Pereira, M. A. Piteri, A. N. Souza, J. P. Papa and H. Adeli, FEMA: A finite element machine for fast learning, *Neural Comput. Appl.* **32**(10) (2020) 6393–6404.
65. Z. Cai, L. Wang, M. Guo, G. Xu, L. Guo and Y. Li, From intricacy to conciseness: A progressive transfer strategy for EEG-based cross-subject emotion recognition, *Int. J. Neural Syst.* **32**(3) (2022) 2250005.
66. B. García-Martínez, A. Fernández-Caballero, A. Martínez-Rodrigo, R. Alcaraz and P. Novais, Evaluation of brain functional connectivity from electroencephalographic signals under different emotional states, *Int. J. Neural Syst.* **32**(10) (2022) 2250026.
67. M. H. Rafiei, L. V. Gauthier, H. Adeli and D. Takabi, Self-supervised learning for electroencephalography, *IEEE Trans. Neural Netw. Learn. Syst.* **35**(2) (2024) 1457–1471.